Predictive Modeling

Week 4

ACSC412/512

Fall 2017

Richard Xu

What were covered last time

1. One GLM example
2. How to assess GLM model (deviance, AIC/BIC, etc.)
3. Claim count/frequency/Cost model
4. Offset and Weights

What we are going to discuss today

1. Exam I, and hw 2
2. Claim count/frequency/Cost model (continued)
3. Classification model
4. Brief summary of GLM
5. Modeling in the real world (specification, variable selection, validation, results visualization, process)
6. Project

Binary Data

Two forms

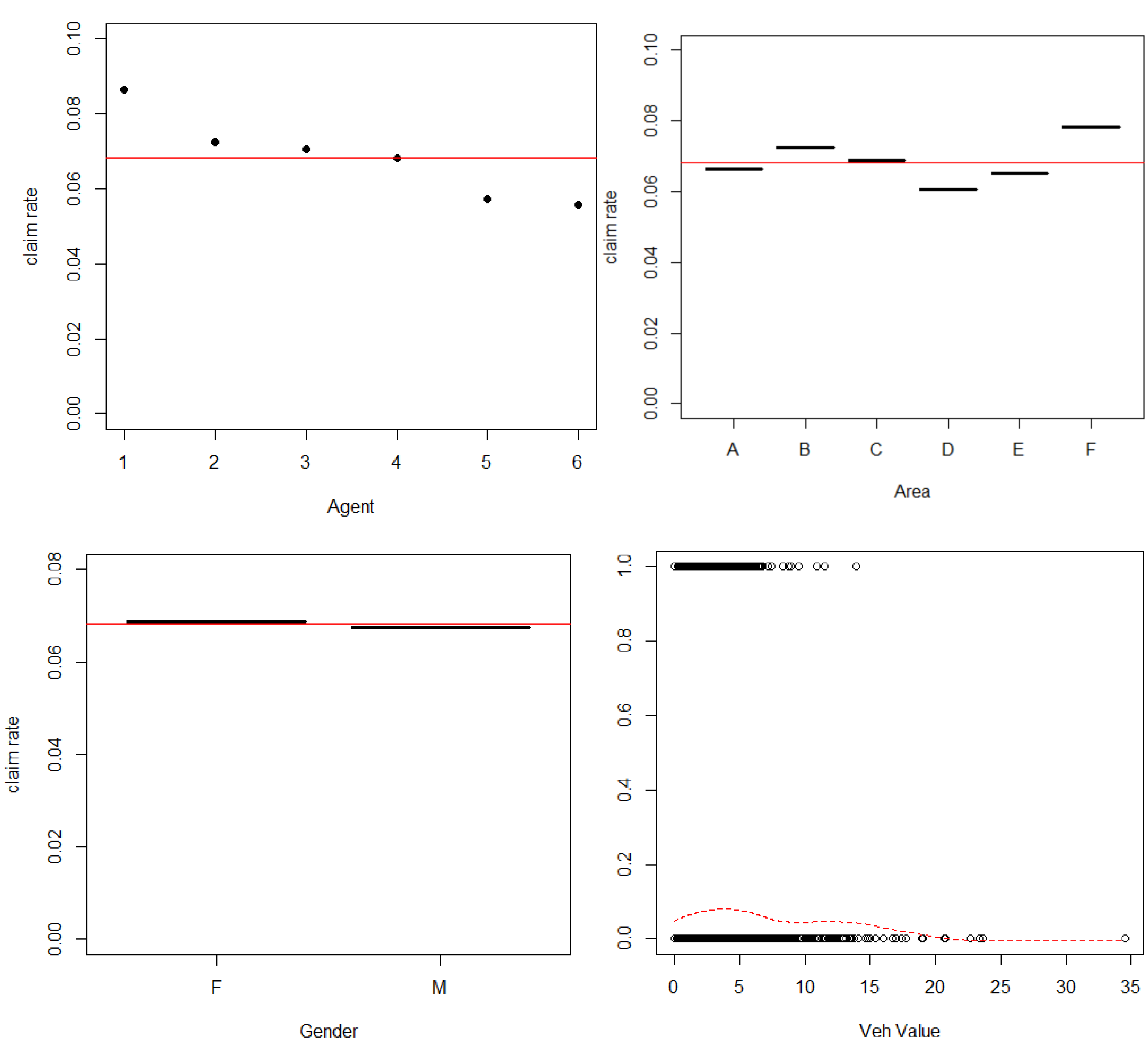
- Ungrouped, variable can take one of two values, say standard/non-standard, or lapse/in-force
- Grouped, variable is number of successes in given trail or group

Natural distribution is the Binomial(n; p) distribution; special case n = 1 in ungrouped; “grouped” could mean “highly aggregated”

Exploring Binary Data

Explore relationship between response & explanatory variables

- For categorical variable, calculate proportions in each subgroups
- For numeric variable, plot of density can be helpful
- Uni-variate analysis (?!)


claim rate
0.00
0.02
0.04
0.06
0.08
0.10
1
2
3
4
5
6
Agent
claim rate
0.00
0.02
0.04
0.06
0.08
0.10
A
B
C
D
E
F
Area
claim rate
0.00
0.02
0.04
0.06
0.08
F
M
Gender
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35
Veh Value

Default link function in R

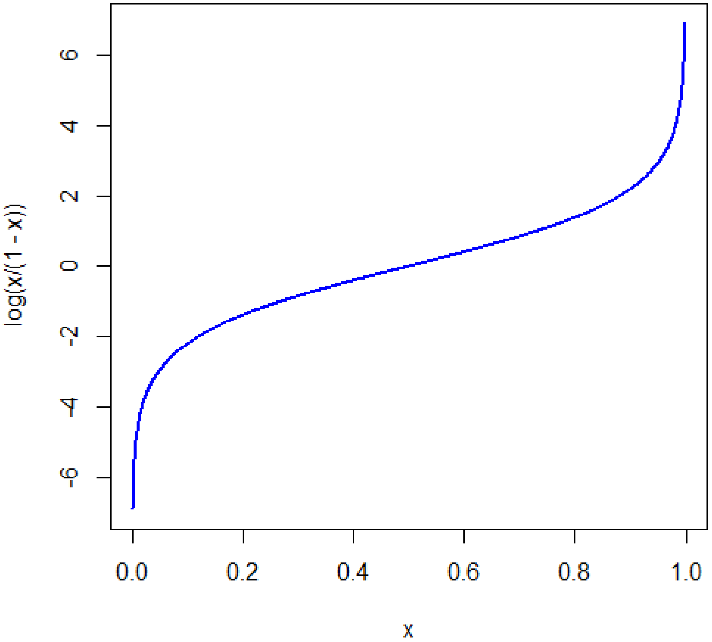
binomial(link = "logit")

Logistics function $f(y) = \frac{1}{1+e^{-y}} = \frac{e^{y}}{1+e^{y}}$

or $\log\left(\frac{p}{1-p}\right) = \eta = \beta X$

p/(1-p) is called odd ratio

$(\log\left(\frac{p}{1-p}\right) \sim p)$



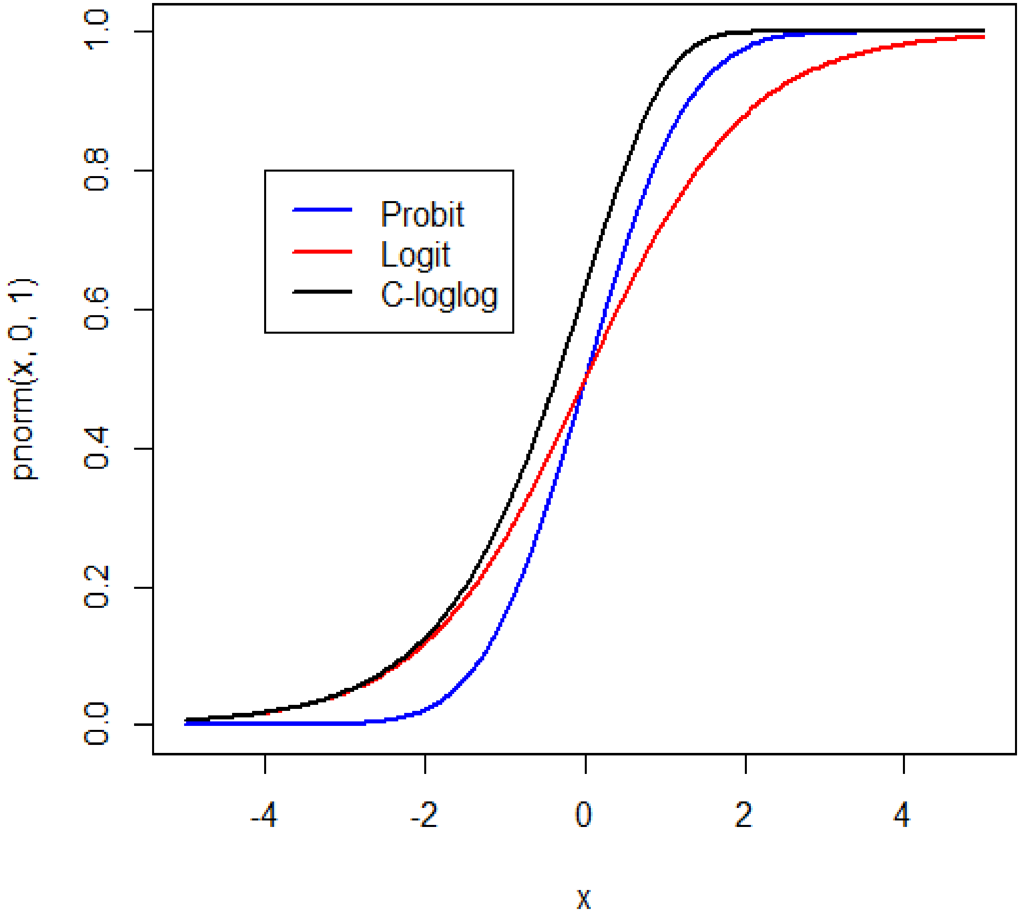
Other link functions

probit – cumulative normal distribution $g(\mu_i) = \Phi^{-1}(\mu_i)$

complementary log-log link - $g(\mu_i) = \log(-\log(1-\mu_i))$

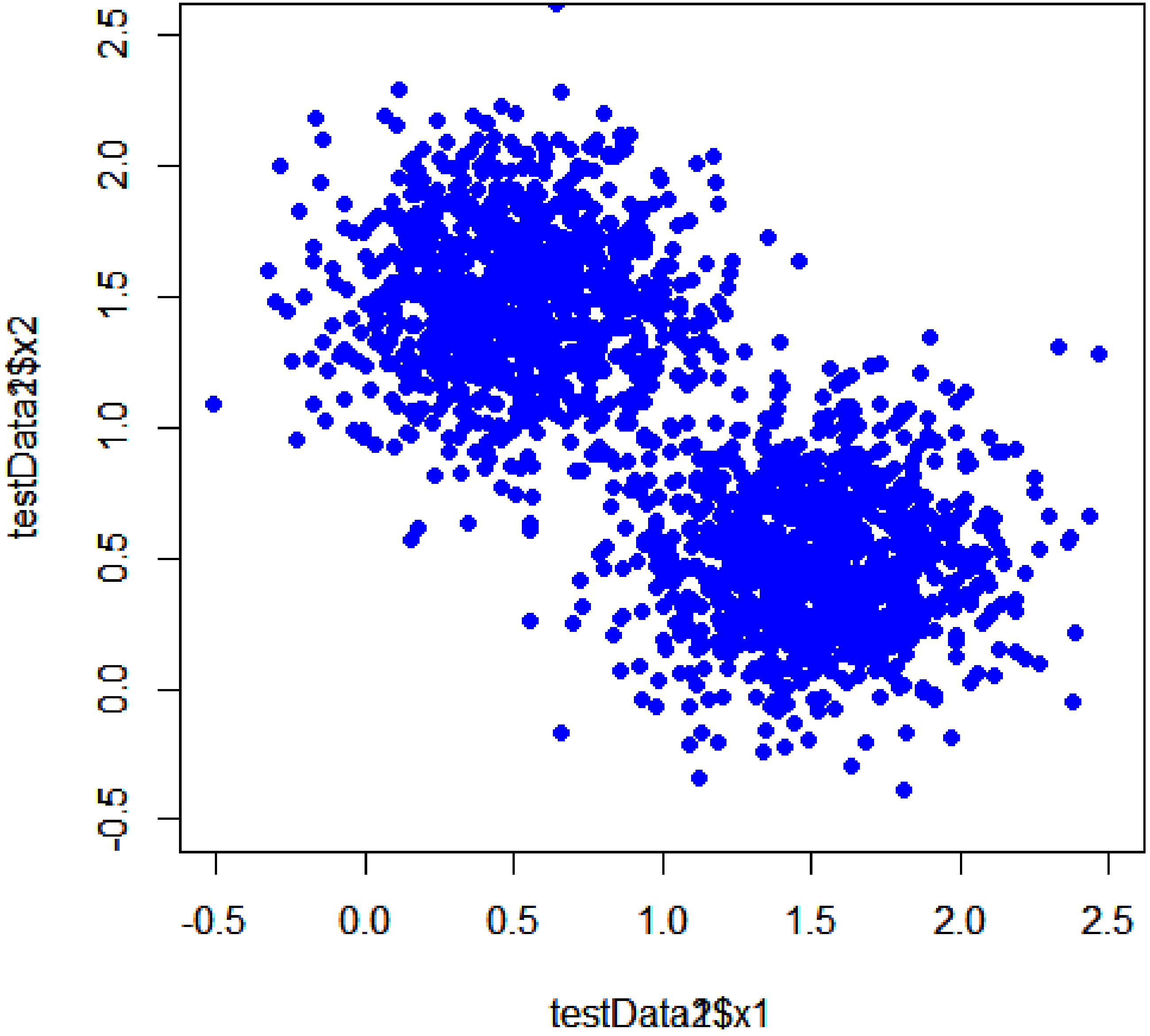
(plot in R, compare to logistics)

What difference and how to choose?

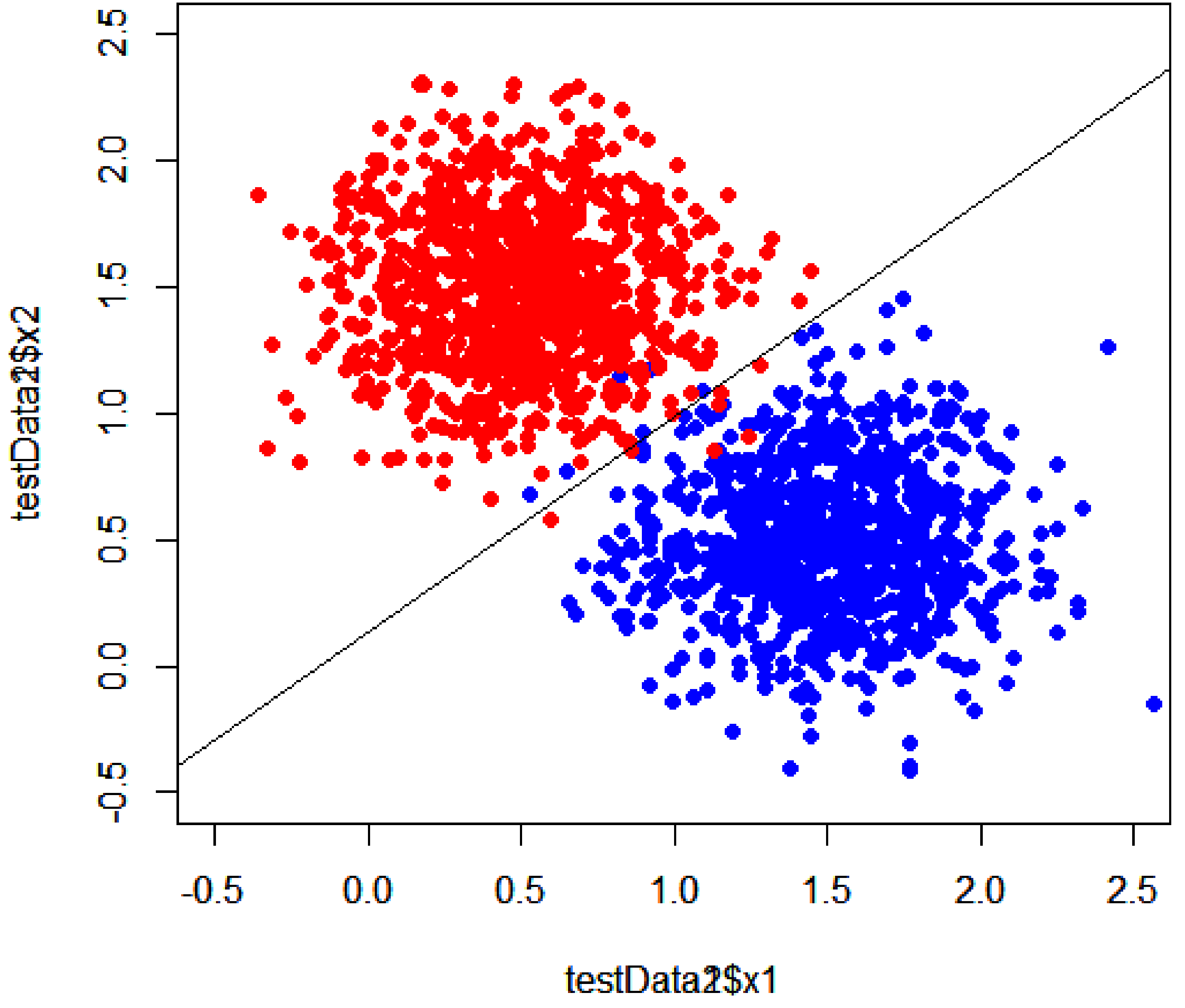


- logit & probit: symmetric & very similar
- logit is preferred due to easy interpretation as logarithm of odds ratio
- Complementary log-log: asymmetric; may be useful when logit inappropriate.

Understand logistic regression



Understand logistic regression



For glm in R, binomial responses can be specified

- A numeric 0/1 vector (0 = death/lapse/disabled, etc.), a logical vector (FALSE = death/lapse/disabled, etc.), or a factor (1st level = death/lapse/disabled, etc.)
- A two-column matrix with numbers of survivals & deaths

```
y <- cbind(death, n - death)
myM1 <- glm(y ~ x1+x3, family = binomial)
```

However, insurance data usually have partial exposure

Problem: warning if partial exposure as inputs in R, as R expects integer as exposure

$$B(1, t\pi), \qquad t \sim exposure, \qquad \pi^* = t\pi$$

$$ln\frac{\pi^*/t}{1-\pi^*/t} = \eta = \beta X$$

$$\pi^* = t\,\frac{1}{1+e^{-y}}$$

So, the link function is not exactly logistic function

(See R script to modify link function.)

Interpretation of model result

$$ln\frac{p}{1-p} = \eta = \beta X$$

Categorical x_i & x_j that are exclusive to each other (e.g. gender)

$$ln\frac{p_i}{1-p_i} = \cdots + \beta_i x_i + \beta_j * 0 + \cdots$$

$$ln\frac{p_j}{1-p_j} = \cdots + \beta_i * 0 + \beta_j x_j + \cdots$$

$$ln\frac{p_i}{1-p_i} - ln\frac{p_j}{1-p_j} = \beta_i x_i - \beta_j x_j = \beta_i - \beta_j$$

Or $\frac{p_i}{1-p_i} / \frac{p_j}{1-p_j} = \exp(\beta_i - \beta_j)$

Interpretation of model result

$$\mathrm{p} = \frac{1}{1+e^{-\eta}} = \frac{e^{\eta}}{1+e^{\eta}}$$

When p is small, e^{y} will be small

$$\mathrm{p} \sim e^{\eta} \sim small$$

$$\eta = \beta X = ln\tfrac{p}{1-p} \sim \ln(p)$$

More like a Poisson distribution, but only when p ~ 0.

(A logistic model to work on)

GLM Summary

	Random	Systematic	Link
OLS	Normal only	$\eta_i = \sum_j x_{ij}\beta_j$	$E(y_i) = \eta_i$
GLM	Various distributions		$g(E(y_i)) = \eta_i$

Great flexibility

- Various distributions & variance structures
- Prior weight & credibility of data
- Offset of data to deal with known effect
- Base levels and intercept

Easy to understand & communicate

Multiplicative model intuitive & consistent

Model Specification

Select model & data distribution

Variable Selection - Stepwise procedure

- **Backward deletion**
  Start with all variables, test the deletion of each variable by using model selection criterion, delete the variable (if any) that improves model the least, repeat this process until no further possible deletion
  (“drop1” very useful in R)
- **Forward addition**
  Start with constant term only, test the addition of each variable by using model selection criterion, add the variable (if any) that improves the model the most, repeat this process until none improves the model
  (“add1” in R)

- **Combination of inclusion & exclusion**
  Combination of the above, test at each step for variables to be included or excluded, loop over all variables and iterate a few times (combine add1 & drop1 in R)
- **Business-Orientated approach**
  Start with variables that have significant business meaning, test all other variables to be included or excluded; most appropriate for business that we know very well

Each approach has its pros and cons.

Can you name a few?
Consider different scenarios
many variables vs. small number of variables

Other major modeling issues

- **Grouping**
  - Problem – too many categories in a variable
  - How to group levels into fewer levels in R
  - Statistics consideration - significance
  - Business consideration – intuitive & manageable
  - Balance between complexity & accuracy (art & science!)

- **Missing value**
  - Welcome to the real world !
  - Fill with median/mode
  - Generate a separate category
  - Imputation of missing value - use correlation between variables to fill in, e.g. package “amelia”

Model Validation & Performance

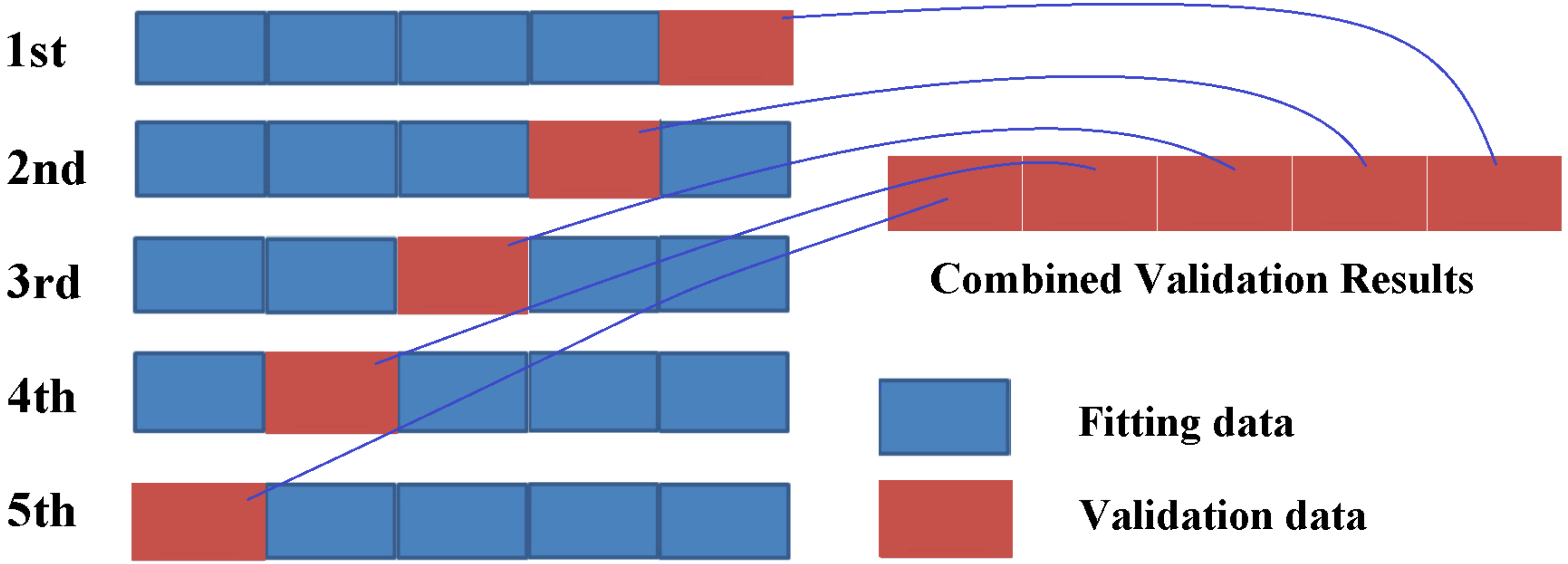
- Can you validate a model?
  - Why can not we use deviance/AIC/BIC? Or in-sample results?
- What is the expected performance in real world?
- Ultimate test of model is actual business
- Goal: model validation and performance expectation

Holdout validation

- Split data into training dataset & validation dataset
  - Training data for fitting & testing model
    - May need to further split into fitting and testing
  - Validation data for assessment of model
  - Usually 50-80% for training/testing, rest for validation
- Intuitive and easy to understanding, but need large enough quantity of data

K-fold Cross-validation

- Dataset: split into k equal size sun-datasets
- Use (k-1) sub-datasets to fit model, the k-th for validation
- Rotate & loop over all sub-datasets; combine results
- In most cases, k=10 is sufficient; K=2 is a special case
- Example k=5



Other Cross-validation

- Leave-one-out validation
  - K=number of records
- Leave-p-out validation
  - loop over all combinations of p out of n
- Repeated random sampling CV

Pros/Cons

- Efficient use of data
- Cost in computing
- Sometimes hard to explain

(look at an example)

Result Presentation

Some considerations

- **Understand your audiences** - know their background & level of knowledge about subject & modeling, and properly set level of technical sophistication
- **Avoid technical jargon** - most actuaries do not have a deep understanding of statistical modeling; avoid terms such as AIC/BIC, deviance, etc. If technical detail is needed, put in appendix
- **Visualize results** - A display of model output will not convey much useful information. Plots and tables are always better.
- **Adhere to actuarial standard** - modeling process is part of actuarial work. Any rule (e.g. ASOPs) that would apply to actuarial work will be applicable to modeling as well.

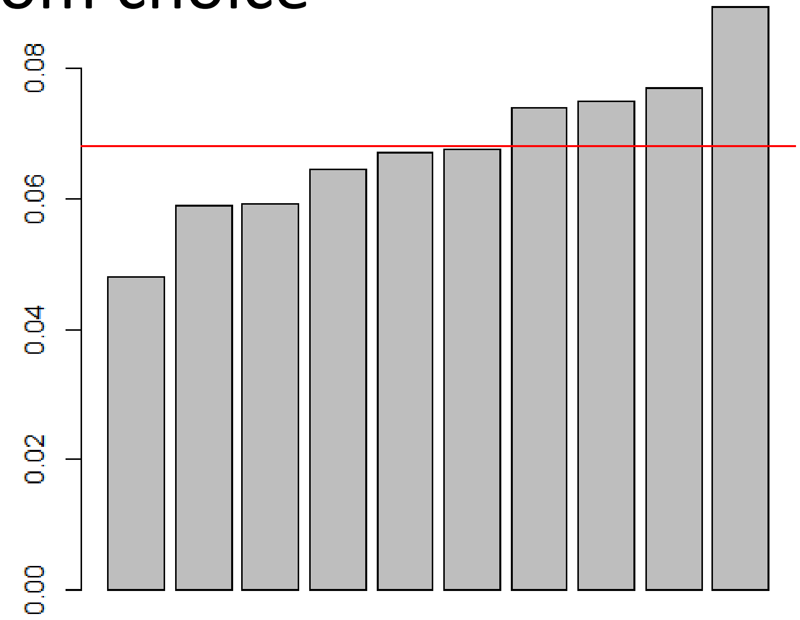
Lift Curves

Measure the performance of a model based on how much improvement compared to random choice

- Model outputs (validation results) are sorted & partitioned into ***n*** equal size blocks
- Analyze the target variable to evaluate performance
- Compare to results from random choice

n =10, each group called a decile

Also, *n* = 4 (quartile) or n=20 also common

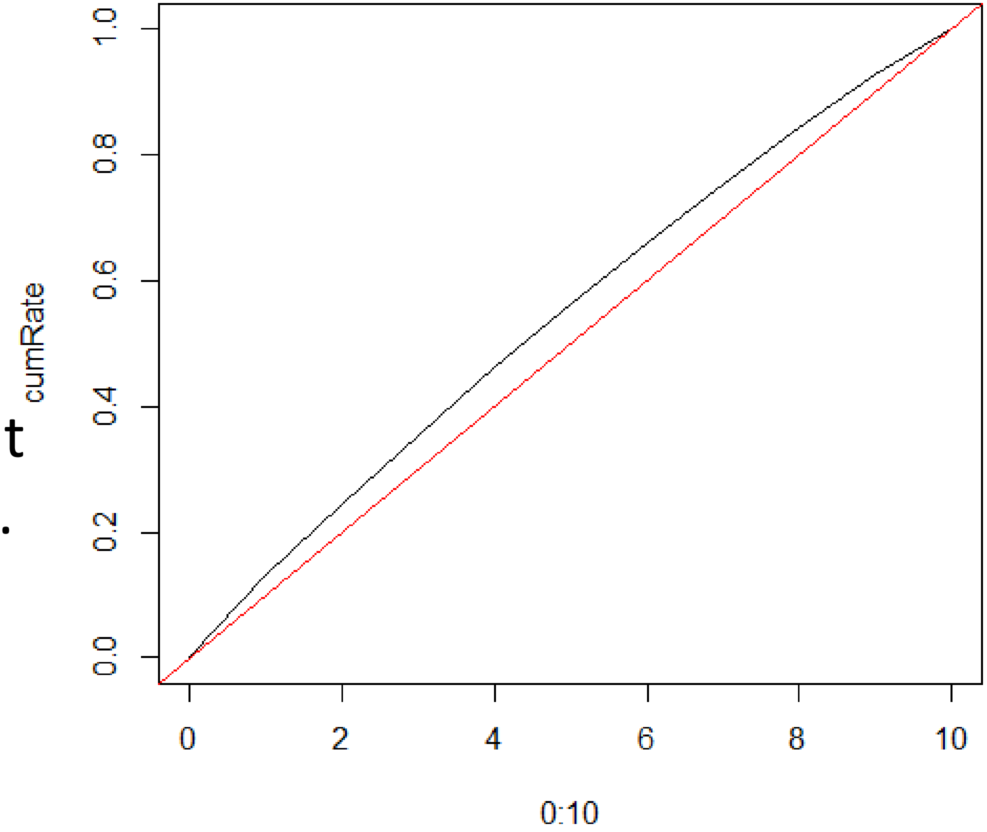


Gain Curve

Cumulative distribution of targeted variable, vs. random selected reference line that is a straight diagonal line from (0,0) to (1,1)
A perfect gain curve is located to the northwest corner (0,1)

Similar to ROC (receiver operating characteristic)

- Gain curve: how a model can differentiate signal
- ROC: what gain (true positive) can be obtained at certain cost (false positive).
- Nearly identical when fraction of background or signal is very small,

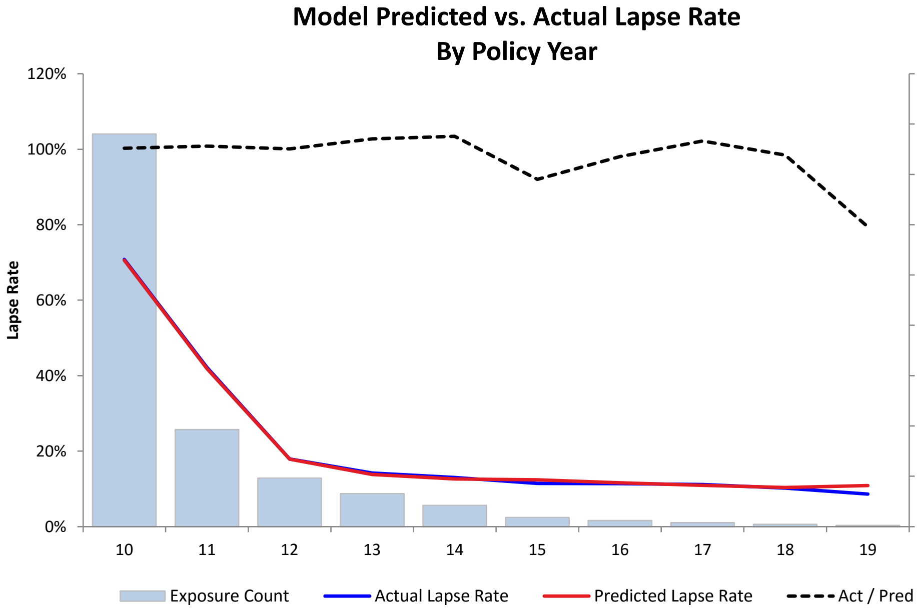


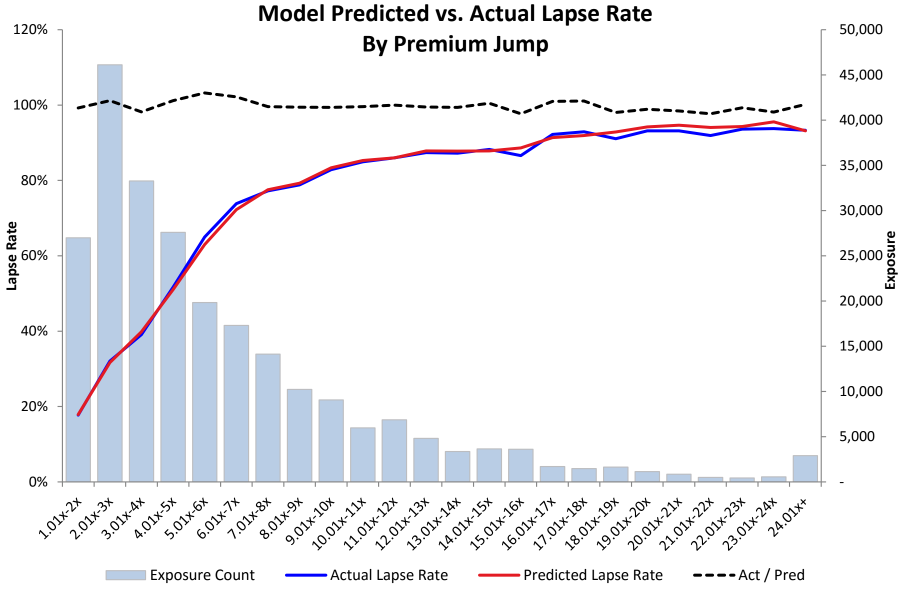
Actual to Expected (A/E)

MARYVILLE UNIVERSITY

More traditional actuarial approach, straightforward if comparing model predicted values to observed data actuaries can easily understand the model's predictive power.

Inclusion of exposure can help to understand how credible the data are.

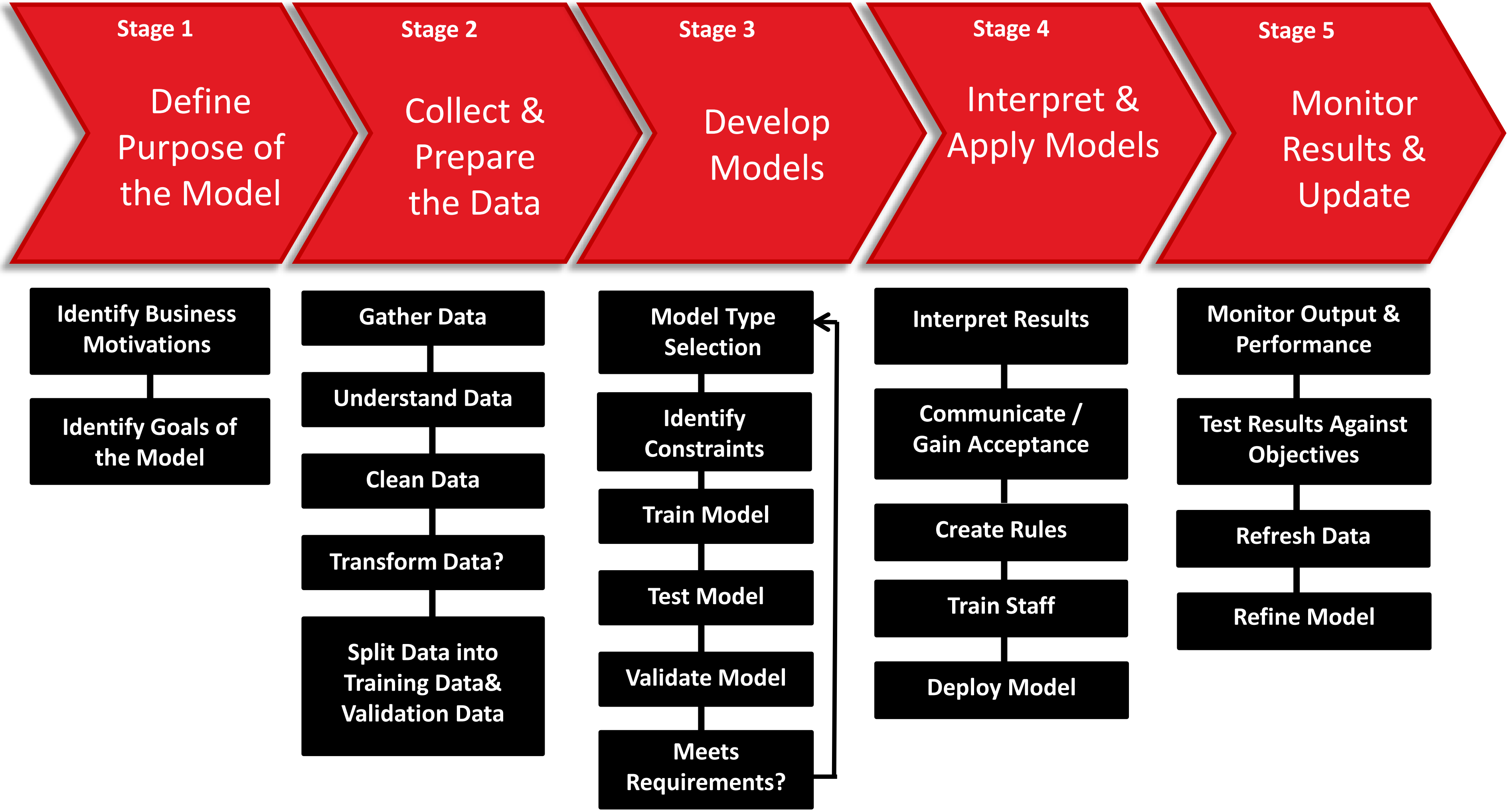

Model Predicted vs. Actual Lapse Rate
By Policy Year


Model Predicted vs. Actual Lapse Rate
By Premium Jump

Model Parameter						Validation Results			
Variable		Type	Coefficient	P-value	Factor	%	Actual	Predicted	A/E
Intercept		-	3.246	2.03E-14					
Issue Age	Issue Age	Numerical	1.621E-01	<2.00E-16					
	(Issue Age)^2	Numerical	-6.419E-04	<2.00E-16					
	log(Issue Age)	Numerical	-2.725	<2.00E-16					
Risk Class	Super-Pref NS	Categorical	0		1.00	17.0%	82.4%	82.5%	100%
	NS		0.03427	1.59E-09	1.03	70.5%	68.7%	68.2%	101%
	SM		0.1205	<2.00E-16	1.13	12.5%	67.4%	68.3%	99%
Face Amount	<50K	Categorical	0		1.00	0.3%	49.5%	55.7%	89%
	50-100K		0.3153	3.49E-15	1.37	6.4%	63.4%	63.0%	101%
	100K-250K		0.3437	<2.00E-16	1.41	43.9%	69.2%	68.9%	100%
	250K-1M		0.3652	<2.00E-16	1.44	41.2%	72.3%	72.1%	100%
	>1M		0.3645	<2.00E-16	1.44	8.2%	79.0%	79.3%	100%
Premium Mode	Annual	Categorical	0		1.00	22.8%	85.5%	85.0%	101%
	Semi/Quarter		-0.03244	1.16E-11	0.97	39.8%	76.1%	75.8%	100%
	Monthly/BiWeekly		-0.2755	<2.00E-16	0.76	34.4%	53.3%	53.5%	100%
	Other/Unknown		0.02057	0.0586	1.02	3.0%	91.1%	90.5%	101%
Premium Jump	PREM_JUMP 1-2	Categorical	0		1.00	5.6%	27.3%	26.9%	102%
	PREM_JUMP 2-3		1.135	<2.00E-16	3.11	13.8%	42.8%	43.0%	99%
	PREM_JUMP 3-4		1.492	<2.00E-16	4.45	10.5%	52.9%	52.6%	100%
	PREM_JUMP 4-5		1.826	<2.00E-16	6.21	10.5%	65.7%	65.1%	101%
	PREM_JUMP 5-6		2.082	<2.00E-16	8.02	9.0%	76.7%	76.4%	100%
	PREM_JUMP 6-7		2.118	<2.00E-16	8.31	8.8%	82.3%	81.7%	101%
	PREM_JUMP 7-8		2.176	<2.00E-16	8.81	7.5%	84.0%	84.1%	100%
	PREM_JUMP8-10		2.246	<2.00E-16	9.45	10.6%	86.2%	85.9%	100%
	PREM_JUMP 10-12		2.304	<2.00E-16	10.01	7.6%	89.0%	88.5%	101%
	PREM_JUMP 12-16		2.342	<2.00E-16	10.40	9.1%	89.4%	89.7%	100%
	PREM_JUMP 16-20		2.385	<2.00E-16	10.86	3.6%	92.8%	92.8%	100%
	PREM_JUMP >20		2.356	<2.00E-16	10.55	3.4%	93.7%	93.9%	100%
Cross Term	Issue Age & PREM_JUMP	Mixed							

Predictive Modeling Process – High Level View

MARYVILLE UNIVERSITY



Stage 1

Define Purpose of the Model

- Identify Business Motivations
- Identify Goals of the Model

Stage 2

Collect & Prepare the Data

- Gather Data
- Understand Data
- Clean Data
- Transform Data?
- Split Data into Training Data& Validation Data

Stage 3

Develop Models

- Model Type Selection
- Identify Constraints
- Train Model
- Test Model
- Validate Model
- Meets Requirements?

Stage 4

Interpret & Apply Models

- Interpret Results
- Communicate / Gain Acceptance
- Create Rules
- Train Staff
- Deploy Model

Stage 5

Monitor Results & Update

- Monitor Output & Performance
- Test Results Against Objectives
- Refresh Data
- Refine Model

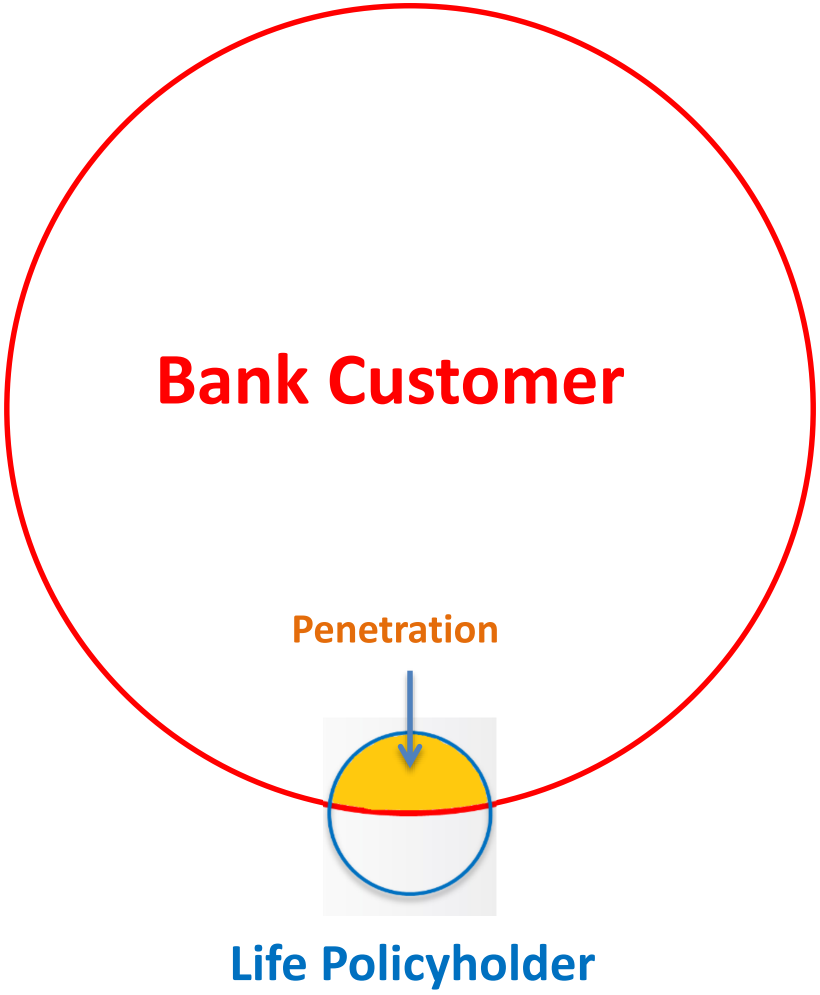
Project

Real life project!

- Background Introduction
- Objectives
- Data
- Modeling
- Final deliverables

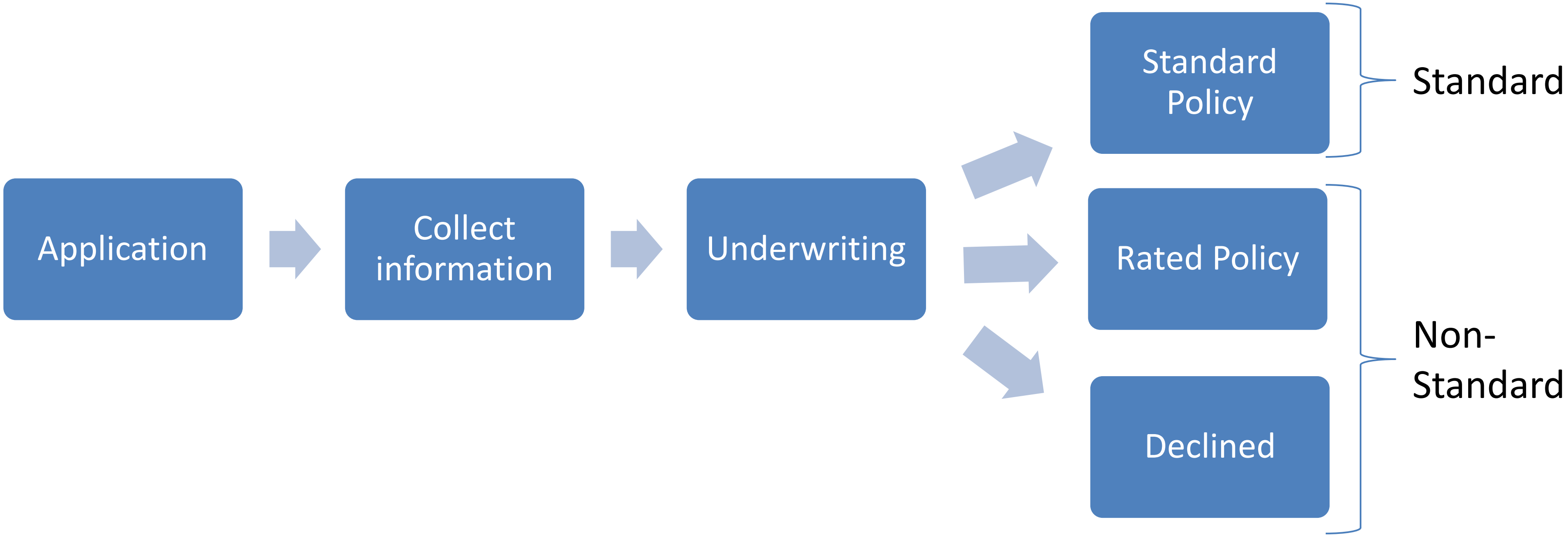
Introduction

- Very large pool in bank customers ~ millions
- Life insurance product – whole life
- UW – medical, but only standard and rated
- Low penetration in life product ~ few %
- Cross-sell life product to existing bank customers
- Major obstacle: medical UW



Goal: Using PM to solve business problem

- Ideally, predict individual mortality & offer product
  - But to do that, what do we need?
- Alternatively, predict UW decisions
  - What are advantage and disadvantage?



Goal: Using PM to solve business problem

- Based on model output, identify pre-qualified low risk customers & offer guaranteed issue (GI) or simplified issue (SI) w/o medical UW
- Business benefit
  - Significant low acquisition costs
  - Short turn-around time & high take rate
  - Deeper market penetration/high sales production
  - Better business performance

Bancassurance is unique for PM

- Insurance data dilemma – enough/not enough data?
- 3rd party data are always required beyond traditional actuarial study
- Financial/demographic information about customers readily available
    - How much does your bank know about you?

Data

- To meet business objective, what are considerations for modeling data?
    - Sources, target variable, predictor variable, timing, etc.

Data

- Find full underwritten life cases that also are bank customer; combine life & bank financial data
- UW decision is the target variable replicated by model
- At the time of UW
- How much bank knows about its customers

Major challenge - limited data

- A total of about <10k full UW cases
- UW decision STD/non-STD w/ very low nonSTD ~3.0%
- Many missing values due to old time, esp. for declined
- Not all information collected at the time of UW

Final deliverables

- Complete model with proofed performance
- A comprehensive document to client; your target audience will be mid-/high-level mangers in bank/life, & actuaries in life product development/pricing/valuation
- A presentation is also required (both groups)

Logistics

- A team of two students randomly assigned (by R, of course!)
- Both are required to make equal contributions to final deliverables; each have 10-15 min presentation, focusing on different aspects of the project
- Project paper is due before the last class
- Dataset & summary will be sent to you this week
- 30-min discussion at the end of each class of week 5/6